ABSTRACT OF THE DISCLOSURE

The invention is a computer-based method for analyzing two versions of an HTML document that identifies new or changed areas of the document while preserving the original textual formatting, including embedded graphics. An HTML document is scanned and the information organized into groupings of HTML tags and text. A set of rules determines which HTML tags are permitted within a group, and which mark the start of a new group. Tags that mark the start of a new group are usually those that break the flow of text when an HTML page is rendered. As a result, the text that constitutes a paragraph, embedded hypertext links, and any associated HTML character-formatting elements are contained within a single group. A modified version of the same HTML document is similarly processed. At this point, the two can be compared group by group in order to detect differences. Any group that does not match the associated group in the original is considered to be a modified group. The modified groups can then be inserted as sections into a new HTML document, and these sections appear to have nearly all of the original formatting intact. Thus, they appear as clipped sections from the original HTML document, and are useful for depicting regions of interest.